The role of artificial intelligence in disinformation

Bontridder, Noemi; Poullet, Yves

Published in:
Data & Policy

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Data & Policy (2021), 3: e32
doi:10.1017/dap.2021.20

RESEARCH ARTICLE

CAMBRIDGE UNIVERSITY PRESS

The role of artificial intelligence in disinformation

Noémi Bontridder* and Yves Poullet

Research Centre in Information, Law and Society, University of Namur, Namur, Belgium
*Corresponding author. E-mail: noemi.bontridder@unamur.be

Received: 02 March 2021; **Revised:** 04 July 2021; **Accepted:** 19 July 2021

Key words: artificial intelligence; content regulation; digital ecosystem; disinformation; online platforms

Abbreviations: AI, artificial intelligence; EU, European Union; ICT, information and communication technologies.

Abstract

Artificial intelligence (AI) systems are playing an overarching role in the disinformation phenomenon our world is currently facing. Such systems boost the problem not only by increasing opportunities to create realistic AI-generated fake content, but also, and essentially, by facilitating the dissemination of disinformation to a targeted audience and at scale by malicious stakeholders. This situation entails multiple ethical and human rights concerns, in particular regarding human dignity, autonomy, democracy, and peace. In reaction, other AI systems are developed to detect and moderate disinformation online. Such systems do not escape from ethical and human rights concerns either, especially regarding freedom of expression and information. Having originally started with ascending co-regulation, the European Union (EU) is now heading toward descending co-regulation of the phenomenon. In particular, the Digital Services Act proposal provides for transparency obligations and external audit for very large online platforms' recommender systems and content moderation. While with this proposal, the Commission focusses on the regulation of content considered as problematic, the EU Parliament and the EU Council call for enhancing access to trustworthy content. In light of our study, we stress that the disinformation problem is mainly caused by the business model of the web that is based on advertising revenues, and that adapting this model would reduce the problem considerably. We also observe that while AI systems are inappropriate to moderate disinformation content online, and even to detect such content, they may be more appropriate to counter the manipulation of the digital ecosystem.

Policy Significance Statement

This study aims at identifying the right approach to tackle the disinformation problem online with due consideration for ethical values, fundamental rights and freedoms, and democracy. While moderating content as such and using AI systems to that end may be particularly problematic regarding freedom of expression and information, we recommend countering the malicious use of technologies online to manipulate individuals. As considering the main cause of the effective manipulation of individuals online is paramount, the business model of the web should be on the radar screen of public regulation more than content moderation. Furthermore, we do support a vibrant, independent, and pluralistic media landscape with investigative journalists following ethical rules.

1. Introduction

Manipulation of truth is a recurring phenomenon throughout history.[1] *Damnatio memoriae*, namely the attempted erasure of people from history, is an example of purposive distortion of reality that was already practiced in Ancient Egypt. Nevertheless, owing to the rapid advances in information and communication technologies (ICT) as well as their increasing pervasiveness, disingenuous information can now be produced easily and in a realistic format, and its dissemination to a targeted audience occurs at an unparalleled speed and scale, including through artificial intelligence (AI) techniques. The consequences are serious with far-reaching implications. For instance, the media ecosystem has been leveraged to influence citizens' opinion and voting decisions related to the 2016 US presidential election[2] and the 2016 UK referendum on leaving the European Union (EU) (Howard and Kollanyi, 2016). In Myanmar, Facebook has been a useful instrument for those seeking to spread hate against Rohingya Muslims (Human Rights Council, 2018, para 74).[3] In India, rumors on WhatsApp resulted in several murders (Dixit and Mac, 2018). In France, a virulent online campaign on social media against a professor ended up with him being murdered (Bindner and Gluck, 2020). Conspiracy theories are currently prospering.[4] And presently in the context of the Covid-19, we are facing what has been called an *infodemic*[5] by the World Health Organization (WHO), with multiple adverse effects on individuals and society at large.

As commonly understood, *disinformation* is false, inaccurate or misleading information that is shared with the intent to deceive the recipient,[6] as opposed to *misinformation* that refers to false, inaccurate, or misleading information that is shared without any intent to deceive. Whereas new digital technology and social media have amplified the creation and spread of both mis- and disinformation, only disinformation has been considered by the EU institutions as a threat that must be tackled by legislative and technical means.[7] This choice of focus has to do with the manipulative character of disinformation, along with the importance of protecting fundamental rights and freedoms, especially freedom of expression and information.[8] Indeed, if anyone or any entity was allowed to decide whose

[1] For an historical review of the phenomenon in the EU, see Mork (2020).

[2] For an in-depth study on the disinformation phenomenon in the US, see Lance Bennett and Livingston (2020). During the 2016 US election campaign, not only has Cambridge Analytica played an important role in influencing voter's decisions, but also the writing of sensationalist stories to earn money from advertising. See for instance the example of the "fake news" farm in Macedonia, reported by Kirby (2016).

[3] The report indicates that "[t]he extent to which Facebook posts and messages have led to real-world discrimination and violence must be independently and thoroughly examined."

As a contributing vector of the possibly large extent to which Facebook was used to spread hate effectively, it is important to note that in Myanmar, Facebook is considered as the internet for most people. This situation is the result of various factors: the social network is already installed when buying a phone; in 2016 it released its "Free Basics" program that offers users creating a Facebook account free access to some pre-selected websites, with the aim to conquer the developing countries; and since people in Myanmar have been silenced for long, they may have particularly reveled in being able to express themselves and share information on social media. These elements were discussed in an article of The Economist (2020a). See also the analysis of the situation conducted by the anthropologist Christina Fink, such as exposed in her article, Fink (2018).

[4] QAnon is an example of such conspiracy theory, see Hannah (2021).

[5] As defined by the WHO, "infodemics are an excessive amount of information about a problem, which makes it difficult to identify a solution. They can spread misinformation, disinformation, and rumors during a health emergency. Infodemics can hamper an effective public health response and create confusion and distrust among people." WHO (2020a). See also WHO (2020b).

[6] The European Commission defines disinformation as "verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public, and may cause public harm. Public harm comprises threats to democratic political and policy-making processes as well as public goods such as the protection of EU citizens' health, the environment or security. Disinformation does not include reporting errors, satire and parody, or clearly identified partisan news and commentary." European Commission (2018d, pp. 3–4). By disseminating disinformation online, malicious stakeholders may for instance seek to discredit political leaders, tarnish the reputation of a competing firm, or negate facts of public interest with antidemocratic purposes.

[7] See for instance European Commission (2018b). It will be revised and strengthened as announced in the European Democracy Action Plan, which addresses the fight against disinformation as one of its three pillars. See European Commission (2020b).

[8] Freedom of expression and freedom to receive information are set out in article 19 of the Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR). They are protected in Europe by article 10 of the Convention for the

truth should be considered as false and was enabled to regulate content accordingly, freedom of expression and information would be seriously impaired. The disinformation problem is particular in the sense that, firstly, the shared information is intentionally deceptive to manipulate people and, secondly, for achieving his or her goal, its author takes benefit from the modern techniques of communication and information. For these reasons, our analysis stays on the beaten path, hence the title of this article referring solely to the disinformation problem. It is also worth specifying that unlike "fake news," a term that has been used by politicians and their supporters to dismiss coverage that they find disagreeable, the disinformation problem encompasses various fabricated information and practices going beyond anything resembling "news."[9]

As indicated, advances in ICT have changed the way information can be produced and disseminated. What must be noted is the decisive role of AI techniques used in this field. Not only do they facilitate the creation and dissemination of disinformation by malicious stakeholders, they are also used contrariwise to tackle disinformation online. In the present study, we first analyse the different AI techniques that amplify the disinformation problem. Second, we focus on AI techniques developed in response to this exact same issue. Ethical implications arise in both cases, which we consider respectively. Third, we discuss the EU regulation of the phenomenon, which started with ascending co-regulation but is presently heading toward descending co-regulation. And finally, we conclude our study and recommend future directions to address the problem ethically, with due consideration for fundamental rights and freedoms.

2. AI Techniques Boost the Creation and Dissemination of Disinformation

AI techniques boost the disinformation phenomenon online in two ways. First, AI techniques are generating new opportunities to create or manipulate texts and image, audio or video content. Second, AI systems developed and deployed by online platforms to enhance their users' engagement significantly contribute to the effective and rapid dissemination of disinformation online. These latter techniques constitute the main contributing factor of the problem. Multiple ethical implications arise from this situation, which should be thoroughly examined.

2.1. AI techniques facilitate the creation of fake content

When AI techniques are used to create fake content, the product is called a *deepfake.* As highlighted in a report dealing with technology-enabled disinformation, "[f]alse media has existed for as long as there has been media to falsify: forgers have faked documents or works of art, teenagers have faked driver's licenses, etc. With the advent of digital media, the problem has been amplified, with tools like Photoshop making it easy for relatively unskilled actors to perform sophisticated alterations to photographs" (Akers et al., 2018, p. 4). More recently, developments in AI have further expanded the possibilities to manipulate texts, images, audios and videos, with the two latter types of content becoming increasingly realistic. The following definition explains clearly what deepfakes are:

> Deepfakes (a portmanteau of deep learning and fake) are the product of two AI algorithms working together in a so-called Generative Adversarial Network (GAN). GANs are best described as a way to algorithmically generate new types of data from existing datasets. For example, a GAN could analyse thousands of pictures of Donald Trump and then generate a new picture that is similar to the analysed images but not an exact copy of any of them. This technology can be applied to various types of content—images, moving images, sound, and text. The term deepfake is primarily used for audio and video content (Walorska, 2020, p. 9).

Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR) and in the EU by article 11 of the Charter of Fundamental Rights of the European Union (CFR).

[9] These are the two reasons why the European Commission has decided to avoid the term "fake news." See European Commission (2018a, p. 10).

While the manipulation of texts and images was already feasible with less sophisticated tools, the advent of deepfakes renders the creation of fake videos and audios highly accessible as well. Indeed, "only a few hundred pictures or audio recordings are required as training data to achieve credible results. For just under $3, anybody can order a fake video of a person of their choice, provided that they have at least 250 pictures of that person—but this is unlikely to be an obstacle for any person that uses Instagram or Facebook. Synthetic voice recordings can also be generated for just $10 per 50 words" (Walorska, 2020, p. 15). Apart from their entertainment value, the involvement of deepfakes in the media can adversely affect society. Not only do they fuel the spread of false information, they are also prone to undermine the credibility of legitimate information, creating doubts about any information encountered, including when it is given by the traditional press or by governmental administrations. In the disinformation context, it is therefore possible for anyone willing to deceive or mislead individuals, to manipulate the truth in two effective ways: fake content can be passed off as real and authentic information can be passed off as fake.

2.2. AI techniques present on the web boost the dissemination of disinformation

To ensure dissemination of disinformation to a maximum of selected people, *micro-targeting* is especially effective. Owing to the economic model of the web based on advertising, tracking methods using AI technologies were developed by information and communication platforms to enable the targeting of advertisements to specific consumers for the sake of efficiency. The traditional *browser cookie-based tracking* or *third-party tracking* "is the practice by which companies embed content—like advertising networks, social media widgets, and website analytics scripts—in the first party sites that users visit directly. This embedding relationship allows the third party to re-identify users across different websites and build a browsing history" (Akers et al., 2018, p. 4). *Browser fingerprinting* is another tracking method, "which relies on browser-specific and OS-specific features and quirks to get a fingerprint of a user's browser that can be used to correlate the user's visits across websites" (Akers et al., 2018, p. 4). As just said, these methods were developed to target advertisements at potential consumers. In this respect, "[t]he advertiser usually sets the targeting parameters (such as demographics and presumed interests), but the platform's algorithmic systems pick the specific individuals who will see the ad and determine the ad's placement within the platform" (Maréchal and Biddle, 2020, p. 13).

Tracking methods are now often used to, more broadly, target particular content at each user. Indeed, according to the economic model adopted by the major online platforms and websites, which is based on "the economics of attention" (Festré and Garrouste, 2015), the financing of investments is not done directly by visitors but rather through the remuneration received from the advertising companies.[10] Therefore, more time spent by a user on a platform means more economic gain for the latter. Algorithms are thus recording every action we take online (which can be active or passive) in order to propose content that will optimize the time we spend using the platform and to propose products we are most likely to purchase. Therefore, as outlined by the Council of the EU in its recent conclusions on safeguarding a free and pluralistic media system, algorithms are "affecting the results that users are actively searching for (findability) and the media content that users are passively exposed to (discoverability)" (Council of the European Union, 2020, para 21). This is the business model of the web we are usually not completely aware of when surfing online.

Since data analysis and algorithms are having an increasing impact on the information shown to individuals, each of us going online sees a different version of reality. Facebook's News feed, Twitter's Timeline, and YouTube's recommender system are only some examples of content shaping algorithms that determine what individual users see online (Maréchal and Biddle, 2020, p. 13). We can also mention

[10] For instance, in 2020, Google's advertising revenue amounted to 146.92 billion US dollars. See https://www.statista.com/statistics/266249/advertising-revenue-of-google/ (accessed 01 March 2021). And Facebook's advertising revenue amounted to 84.2 billion U.S. dollars. See https://www.statista.com/statistics/267031/facebooks-annual-revenue-by-segment/ (accessed 01 March 2021).

Netflix's movie recommender system, Spotify's recommender system and, more surreptitiously, Google's ranking system. Besides privacy concerns raised by the tracking and targeting technologies as well as concerns regarding people's autonomy and right to information, this ecosystem can be directly leveraged by disinformation campaigns to target specific vulnerable users and to create and exploit filter bubbles (Akers et al., 2018 , p. 5). For instance, there existed until recently on Facebook a pseudoscience interest category that advertisers could purchase and target (Sankin, 2020), and since AI systems analyse the unique psychographic and behavioral user profiles, micro-targeting of voters has been boosted by the use of such technology (Bergamini, 2020, p. 10; Marsden and Meyer, 2019, p. 15).

The design of algorithms based on micro-targeting can also directly amplify the spread of both mis- and disinformation. For example, on YouTube, more than a billion hours of video are viewed every day, 70% of which by automated systems in order to provide recommendations on what video to watch next for human users. Since there are more than two billion users on YouTube, this has a significant impact on what the world watches. Yet, as the platform operates for profit, which is optimized by viewing time, the quality of the recommended content is far from being prioritized. The project AlgoTransparency, which aims at exposing the impact of the most influential algorithms,[11] demonstrates YouTube's vicious feedback loop. It can be described as follows: (1) Divisive content performs better → (2) Algorithmic systems promote this content in order to optimize viewing time → (3) This kind of content being more viewed, content creators create more of it.[12]

The deployment of *social bots* by malicious stakeholders also contributes to the effective dissemination of disinformation. These bots (an abbreviation for *software robots*) are fully or semiautomated user-accounts operating on social media platforms, which are designed for communication and the imitation of human online-behavior. There are different types of social bots, ranging "from very simple bots that automate single elements of the communication process (e.g., liking or sharing), over partially human-steered accounts with automated elements (so-called hybrid bots, or 'cyborgs') to autonomously acting agents equipped with [AI] and learning skills such as Microsofts' Zo1 or Replika.ai" (Assenmacher et al., 2020). These "artificial speakers," which are consistently present on social media,[13] have a real impact: they "foment political strife, skew online discourse, and manipulate the marketplace" (Lamo and Calo, 2018). Indeed, such bots can be designed to post context-relevant content based on the community they are attempting to blend into, and once they have gained a credible profile and seem trustable, they are capable of disseminating disinformation efficiently (Akers et al., 2018, p. 5).[14]

2.3. Ethical implications

Disinformation is well a long-standing problem, but given that AI techniques present in the digital ecosystem create new possibilities to manipulate individuals effectively and at scale, multiple ethical concerns arise or are exacerbated. These should be carefully considered.

A first ethical value that is challenged by the current digital ecosystem is *human dignity.*[15] Following this fundamental value, "human beings are to be understood as ends in themselves and never as a means alone" (EDPS Ethics Advisory Group, 2018). Yet, when algorithms are programmed to adapt what is shown to individuals based on their profile created through datafication[16] in order to optimize

[11] See https://www.algotransparency.org/ (accessed 01 March 2021).

[12] For a presentation of the project AlgoTransparency and its main findings by the founder Guillaume Chaslot, see the following video: *The Toxic Potential of YouTube's Feedback Loop* (CADE Tech Policy Workshop, 17 November 2019). https://www.youtube.com/watch?v=Et2n0J0OeQ8&feature=youtu.be (accessed 01 March 2021).

[13] The average presence of bots was estimated to range between 9 and 15% of all Twitter accounts in 2017, and to amount to approximately 11% of all Facebook accounts in 2019. See Varol et al. (2017) and Zago et al. (2019).

[14] For an analysis of which users may be the target of bots' activities, see Balestrucci (2020).

[15] Human dignity is set out in article 1 of the Universal Declaration of Human Rights and in article 2 of the Charter of Fundamental Rights of the European Union.

[16] Following the definition provided on Wikipedia, "Datafication is a technological trend turning many aspects of our life into data which is subsequently transferred into information realised as a new form of value." https://en.wikipedia.org/wiki/Datafication.

engagement, regardless of the content's quality, those individuals are considered as mere means for economic purposes. Furthermore, those same algorithms can be leveraged by malicious stakeholders to effectively manipulate their opinion and to reach specific goals thereupon. As explained by the EDPS Ethics Advisory Group, "[w]hen individuals are treated not as persons but as mere temporary aggregates of data processed at an industrial scale so as to optimize through algorithmic profiling, administrative, financial, educational, judicial, commercial, and other interactions with them, they are arguably, not fully respected, neither in their dignity nor in their humanity" (EDPS Ethics Advisory Group, 2018, p. 17). In addition, AI techniques present in the digital ecosystem change reality in most cases unbeknownst to the individuals, expanding opportunities for effective manipulation of their opinion. Indeed, targeted individuals are rarely aware of the current digital ecosystem, and they usually think that the (dis) information they see online is objective and universally encountered by other users.

Secondly, the difficulty to access information alongside the pervasiveness of disinformation online drastically impairs the individuals' capacity to make free and informed decisions, which is an essential prerequisite for their *autonomy* (Poullet, 2020a). The value of autonomy "refers to the capacity of individuals to construct their own identity, to determine their own 'good', their own vision of a good life in respect of others' similar capacity, and therefore to contribute fully to collective deliberation" (Free translation; Poullet, 2020b, p. 93). As observed by the European Court of Human Rights in the *Pretty* case,[17] this notion of personal autonomy underlies the right to privacy.[18] As algorithms make use of personal data to determine the content that will be shown to each individual and distort their capacity to decide freely, the right to privacy and data protection[19] may be violated.[20]

Information provides individuals with the capacity to make informed decisions by enabling them to acquaint themselves with facts and societal challenges, and understand those (Hanot and Michel, 2020, p. 162). It is therefore a key element of individuals' autonomy. Yet, when individuals encounter realistic fake content, when they are enclosed unconsciously or consciously in filter bubbles and "echo chambers," and when they are the target of disinformation campaigns that leverage the current digital ecosystem to effectively manipulate their opinion, their access to information is definitely made harder, which impedes or at least limits their right to information. The decrease of the average level of trust in the news worldwide,[21] to which the invasiveness of disinformation online has participated, also contributes to the lack of information that individuals have to deal with.

Since citizens' capacity to make free and informed decisions is impaired by the digital ecosystem, their ability to participate in the democratic discourse is also affected. Yet, when individuals' capacity to participate fully in public debate is impaired, including through the effective manipulation of their opinion and voting decisions, *democracy* is seriously jeopardized. Democracy indeed "implies that people with different views come together to find common solutions through dialogue" (Bergamini, 2020, p. 15).

It is important to note that in order to form their own opinion, individuals need to access diversified information, namely to be informed in a contradictory way. Receiving information only oriented toward one point of view is not enough. Alongside disinformation campaigns, echo chambers therefore play a key role in the threat toward the democratic process. They are generated in consequence of the filter bubbles created by micro-targeting methods, namely content shaping algorithms. An echo chamber can be defined as "an environment in which individuals encounter only beliefs or opinions that coincide with their own,

[17] ECtHR, *Pretty v. United Kingdom* (Judgment) [2002] Application n°2346/02, para 61.

[18] The right to privacy is set out in article 8 of the European Convention on Human Rights and in article 7 of the Charter of Fundamental Rights of the European Union.

[19] The right to protection of personal data is included in the right to privacy, but it is also explicitly set out in article 8 of the Charter of Fundamental Rights of the European Union. In the EU, the GDPR sets out rules aimed at ensuring this right. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L 119/1 (General Data Protection Regulation).

[20] For further developments regarding the right to privacy and data protection as guarantees of the individuals' autonomy, see Poullet (2020b, pp. 114–131).

[21] For an analysis of the current level of trust in the media, see Newman et al. (2020).

so that their existing views are reinforced and alternative ideas are not considered" (Bergamini, 2020, note 18). Individuals are thus led "into a state of intellectual isolation where there is no place for dialogue […] [and,] by prioritizing the news and information which users like, algorithms tend to reinforce their opinions, tastes and habits, and limiting access to diverging views" (Bergamini, 2020, p. 10). As outlined by Vincent De Coorebyter, "if we define *democracy* as a mechanism of construction of compromises taking account of a large diversity of opinions and *demagogy* as a method to flatter predetermined ideas, we can fear that the Internet fosters the second more than the first" (free translation; de Coorebyter, 2020).

In 2001, Tim Berners-Lee had already expressed similar concerns regarding the web's evolution to a journalist, who reported the related part of the interview as follows: "Berners-Lee, standing at a blackboard, draws a graph, as he's prone to do. It arrays social groups by size. Families, workplace groups, schools, towns, companies, the nation, the planet. The Web could in theory make things work smoothly at all of these levels, as well as between them. That, indeed, was the original idea—an organic expanse of collaboration. But the Web can pull the other way. And Berners-Lee worries about whether it will 'allow cranks and nut cases to find in the world 20 or 30 other cranks and nut cases who are absolutely convinced of the same things. Allow them to set up filters around themselves … and develop a pothole of culture out of which they cannot climb.' Will we 'end up with a world which is full of very, very disparate cultures which do not talk to each other?'" (Wright, 2001). In 2019, he added that "we have to make sure that the web is serving humanity. Not just by keeping it free and open, but by making sure that the things that people build in this permissionless space are actually helping democracy" (Perrigo, 2019).

Enclosing individuals in filter bubbles and echo chambers is definitely dangerous for *peace*, since living together requires knowledge of and tolerance for different views and cultures. As individuals increasingly interact solely with groups of people with their own views, sometimes inflated by social bots, and are manipulated with a biased version of reality, their ability to accept the presence of other cultures and to understand them is made harder. This situation nurtures radicalisation, thus a lack of tolerance for each other essential for peace and stability. Mahatma Gandhi, well-known for his philosophy of nonviolence, was giving attention to openness as he said: "I do not want my house to be walled in on all sides and my windows to be stuffed. I want the cultures of all lands to be blown about my house as freely as possible."[22]

3. AI Techniques As a Way to Tackle Disinformation Online

In reaction to the alarming consequences of disinformation online, social media platforms and search engines are increasingly requested to respond to the phenomenon, especially now in the context of the "infodemic."[23] Consequently, various technological methods are being developed. In this regard, AI techniques are explored both to detect false, inaccurate or misleading content, and to regulate such content online. An important point to consider from the outset is the inability, or inappropriateness, of AI systems to differentiate misinformation from disinformation, which is particularly problematic regarding freedom of expression and information, as we explain *infra.*

3.1. AI techniques are developed to detect disinformation online

To detect articles containing false information, which are not necessarily deepfakes, the already mentioned technical report dealing with technology-enabled disinformation presents and analyses four techniques (Akers et al., 2018, p. 6). First, it is possible with machine learning to train *end-to-end models* using labeled data, namely articles containing false information and articles containing accurate information. The system is then able to directly differentiate between these two types of articles. However, besides the large amounts of labeled data necessary for such a model to be operational, which can be tricky to obtain, its output would lack explainability and would be affected by biases in the

[22] Mahatma Gandhi, *Young India*, 1921.

[23] See *infra.* The term "infodemic" is defined in note 5.

datasets. Second, the *detection of factual inaccuracies* may be more effective. It can be done by comparing the content of the article with external evidence, but this task is better performed by human fact checkers because the nuances of natural language are difficult to formalize. It can nonetheless also be done by verifying if a claim is backed up by given sources, or by using unstructured external web information. Third, the *detection of misleading style*—inferring the intent of an article by analyzing its style—can be done by trained human experts or through machine learning. However, the presence or absence of misleading style is not always correlated with the inaccuracy or veracity of the provided information. Fourth, the *analysis of metadata* (e.g., the sharer's profile or attributes …) instead of the article's content as such is also investigated.

As highlighted by Chris Marsden and Trisha Meyer, textual analysis programs trained to identify potential disinformation material are prone to false negatives or positives "due to the difficulty of parsing multiple, complex and possibly conflicting meanings emerging from the text" (Marsden and Meyer, 2019, pp. 1–2). It has also been noted that as of 2018, Facebook's AI systems did not have enough training data to be highly effective outside of English and Portuguese (Marsden and Meyer, 2019, p. 17). Accurate detection of articles containing false, inaccurate or misleading information therefore still requires the intervention of human agents.

To detect deepfakes in particular, different techniques have been envisaged as well. A first one is to develop *technologies capable of distinguishing between fake content and real content.* To this end, it is possible to use algorithms similar to those which generated the deepfake.[24] The use of forensic tools can also be considered. Researchers observed in 2018 that actors in AI-generated videos did not blink due to the lack of faces with closed eyes in most training datasets. Technologies looking for abnormal patterns of eyelid movement were thus developed. However, once this method became public, new deepfake videos were adapted by featuring blinking people (Kertysova, 2018, p. 71; Walorska, 2020, p. 24). A second technique that has been envisaged to detect deepfakes is the *authentication of content before it spreads*: if images, audios, and videos can be digitally labeled at the moment of their creation, this label could be used as a reference to compare with suspected fake content (Kertysova, 2018, p. 71). However, deepfakes could be created with pre-existing unlabelled content, which would make this approach inefficient in our opinion. For the record, the "*authenticated alibi service*" is a third technological approach that would monitor and store all individual's locations, movements, and actions in order to prove where each individual was and what he or she was saying or doing at any given moment (Kertysova, 2018, p. 71). This approach is particularly undesirable as it would generate multiple human rights' violations, including regarding privacy.

Given the lack of efficient means to detect deepfakes and their potential impact on the legitimacy of online information, the Partnership on AI (PAI)[25] created the AI and Media Integrity Steering Committee in late 2019 for developing and advising projects that strengthen mis/disinformation solutions, including detection of manipulated and synthetic content. Its first project is the *Deepfake Detection Challenge*,

[24] For instance, "[u]sing GLTR, a model based on the GPT-2 system […], researchers from the MIT-IBM Watson AI Lab and HarvardNLP investigated whether the same technology used to write independently fabricated articles can be used to recognize text passages that were generated by AI. When a text passage is generated in the test application, its words are highlighted in green, yellow, red, or purple to indicate their predictability, in decreasing order. The higher the proportion of words with low predictability, namely sections marked in red and purple, the greater the likelihood that the passage was written by a human author. The more predictable the words (and the 'greener' the text), the more likely the text was automatically generated." Walorska (2020, p. 24).

[25] The Partnership on AI "was formally established in late 2016, led by a group of AI researchers representing six of the world's largest technology companies: Apple, Amazon, DeepMind and Google, Facebook, IBM, and Microsoft. In 2017 the addition of six not-for-profit Board members expanded the Partnership into a multi-stakeholder organization—which now represents a community of 50+ member organizations." This multistakeholder organization now "brings together academics, researchers, civil society organizations, companies building and utilizing AI technology, and other groups working to better understand AI's impacts. The Partnership was established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society." See https://www.partnershiponai.org/ (accessed 01 March 2021).

which aims at promoting the development of technical solutions that detect AI-generated and manipulated videos (for more information, see, Partnership on AI, 2020).

To *detect social bots*, various efforts have been made as well. As exposed in the report dealing with technology-enabled disinformation, "[d]etection approaches commonly leverage machine learning to find differences between human users and bot accounts. Detection models make use of the social network graph structure, account data and posting metrics, as well as natural language processing techniques to analyze the text content from profiles. Crowdsourcing techniques have also been attempted. From the platforms provider's vantage point, a promising approach relies on monitoring account behavior such as time spent viewing posts and number of friend requests sent" (Akers et al., 2018, p. 5). It can be observed that AI techniques have been more successful to detect fake accounts, including bot accounts. Indeed, 99.6% of Facebook's fake accounts actioned in the fourth quarter of 2020 were found and flagged by the company before users reported them.[26]

3.2. AI techniques are developed to regulate content online

AI techniques can also be used to aid regulation of content and accounts online. The following methods are discussed in the report prepared by Marsden and Meyer at the request of the STOA (Marsden and Meyer, 2019, pp. 38–41). *Filtering of content* is a measure undertaken by technical providers to prevent the upload or posting of content, and *removal of content* is undertaken upon awareness, request, or order. Filtering or removal of content is a particularly effective method to tackle disinformation, but also the most invasive one as it prevents content sharing. *Blocking of content* is a method by which the user's access to the content is blocked. However, this method can be circumvented by employing a virtual private network (VPN). *(De)prioritization of content* can either be done by the user who opts to see less content related to particular persons or subjects, or by utilizing automated ranking through algorithms and network-based solutions. *Disabling and suspension of accounts* is another method that can be used to tackle disinformation online. Technology providers take these measures when their users abuse the terms of service or do not respect legislation. As is well-known and disputable, Twitter and Facebook have for instance suspended Donald Trump's account for having violated the platforms' terms and conditions.[27]

Other innovative technologies that may prevent disinformation are being developed, such as *decentralized web (Dweb) technologies.* Indeed, Dweb technologies "would enable us to break down the immense databases that are currently held centrally by internet companies rather than users (hence the decentralized web). In principle, this would also better protect users from private and government surveillance as data would no longer be stored in a way that was easy for third parties to access. This actually harks back to the original philosophy behind the internet, which was first created to decentralize US communications during the Cold War to make them less vulnerable to attack."[28]

3.3. Ethical implications

Even if the intended purpose behind the development of AI techniques in response to the ambient disinformation may be legitimate, their use does not escape from multiple ethical concerns either. Taking them into account is of paramount importance before deploying such techniques in the digital ecosystem.

[26] Facebook, "Community Standards Enforcement Report," https://transparency.facebook.com/community-standards-enforcement#fake-accounts (accessed 01 March 2021).

[27] Twitter, "Permanent suspension of @realDonaldTrump" (8 January 2021). https://blog.twitter.com/en_us/topics/company/2020/suspension.html (accessed 01 March 2021); Facebook, "Referring Former President Trump's Suspension From Facebook to the Oversight Board" (21 January 2021). https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board/ (accessed 01 March 2021).

[28] Harbinja and Karagiannopoulos (2019); "When we currently access the web, our computers use the HTTP protocol in the form of web addresses to find information stored at a fixed location, usually on a single server. In contrast, the DWeb would find information based on its content, meaning it could be stored in multiple places at once. As a result, this form of the web also involves all computers providing services as well as accessing them, known as peer-to-peer connectivity."

First, the use of AI systems to detect disinformation material requires an agreed formal definition of what disinformation encompasses. Yet, defining the problem is not always straightforward (Hanot and Michel, 2020, pp. 157–161; Poullet, 2020c). and raises questions regarding who is the judge in determining what is legal or illegal and desirable or undesirable in society. Distinguishing satire, propaganda, and hoaxes from disinformation may be problematic, and some definitions include these kinds of content in the disinformation problem. As outlined by the European Court of Human Rights in the leading case *Handyside v. United Kingdom*, freedom of expression "is applicable not only to 'information' or 'ideas' that are favorably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no 'democratic society.' This means, amongst other things that every 'formality,' 'condition,' 'restriction' or 'penalty' imposed in this sphere must be proportionate to the legitimate aim pursued."[29] Furthermore, the scope of the problem is questionable. As a matter of fact, the lack of access to accurate information goes beyond the disinformation problem. Misinformation also affects society as a whole, and most people sharing content are unaware of the initial intention of the source's publisher. However, tackling misinformation would dramatically limit *freedom of expression and information*, as we have already mentioned.

Yet, difficulty particularly arises when it comes to distinguish misinformation from disinformation, as the intent of the sharer needs to be determined to that end. As explained in this section, AI techniques developed to tackle disinformation online detect all false, inaccurate or misleading information with no distinction related to the intent of the sharer. Permitting AI systems to regulate content automatically would therefore seriously affect freedom of expression and information. This is the case even when humans are involved in the process, since they may rely on the output given by the AI system. We can pursue our reflection by questioning the possibility to enable AI systems to assess the malicious intent of the sharer. Well, since the intention of a person creating or spreading content can be unclear, it would be particularly difficult to enable a machine to assess the intent behind a shared content with the degree of certainty that the delicate action of moderation should require.

To express false, inaccurate or misleading information is indeed not *per se* condemnable. What is reprehensible is by no means related to the quality of the shared content but rather the malicious use of technology to expand voluntarily false, inaccurate, or misleading information in order to manipulate people. AI techniques might thus be used precisely to detect this malicious use of technology (e.g., AI-generated content, the use of recommender systems to target content at specific individuals with a malicious intent, the use of social bots) but not to assess the information's quality. The Council of the EU indeed concluded that "with regard to the importance of freedom of speech, states and administrative regulatory authorities as well as private platform providers should abstain from defining quality content or the reliability of content itself."[30]

In addition, Marsden and Meyer underline that "it is not clear how often and under which circumstances *ex ante* filtering or blocking take place on the platforms. Some is machine-driven, but it is unclear how the illegality of the content or its violation with the community guidelines is determined, nor what safeguards are in place to prevent over-censoring of content."[31] Yet, transparency of content regulation is primordial in order to assess its legality, especially regarding freedom of expression. Explainability of any content regulation is therefore required, and the complexity of AI systems' functioning is problematic in this regard, as it may lead to opaque outputs.

The Information for All Programme[32] of UNESCO had already highlighted a decade ago that while technologies "can open channels by which information may be shared and opinions expressed; it can also

[29] ECtHR, *Handyside v. United Kingdom* (Judgment) [1976] Application n°5493/72, para 49.

[30] Council of the European Union (2020, para 39). It adds that "[t]his should not prevent platforms from promoting public communications and announcements in case of crisis or emergency situations."

[31] See Marsden and Meyer (2019, p. 45). In this regard, we specify that algorithms deployed by platforms can be quite explainable (machine learning) or, by contrast, quite inexplainable (so-called "black box AI").

[32] https://en.unesco.org/programme/ifap.

be used to restrict the information available and to identify and interfere with people expressing alternative opinions" (Rundle and Conley, 2007, p. 15). They specified that "the right to freedom of opinion and expression loses value without the ability to communicate one's views to others. ICTs can be used to create a public forum where this communication can take place, or it can restrict expression by placing limits on a person's ability to communicate with others" (Rundle and Conley, 2007, pp. 15–16). Albeit the current digital ecosystem generates the necessity to place some limits on one's opportunities to manipulate individuals, programming algorithms to tackle false, inaccurate or misleading information would limit the individuals' ability to express freely their opinion, which is inadmissible, how factually questionable those opinions may be, regarding the fundamental rights to freedom of expression and information necessary for democracy and without which we would readily evolve toward a totalitarian society resembling to the one described by George Orwell in his well-known dystopian novel.

Furthermore, as we have already mentioned, AI systems trained to detect false, inaccurate, or misleading information are prone to false positives and false negatives. False positives, namely the wrongful detection of false, inaccurate, or misleading content, affect freedom of expression. Indeed, they "could lead to over-censorship of legitimate content that is machine-labeled incorrectly as disinformation" (Marsden and Meyer, 2019, p. 17). On the other hand, false positives and false negatives would both generate discriminations, therefore impacting *equality and nondiscrimination*.[33] These can be generated by bias in the algorithms. For instance, we indicated above that Facebook's algorithms lack training data outside of English and Portuguese. Therefore, we can assume that content shared in any other language is less prone to be detected by AI systems. This would generate an asymmetrical approach to the disinformation problem.

Other concerns emerge when AI techniques are used to automatically regulate content. For instance, the use of content ranking algorithms can pose problems regarding *media pluralism* when it relies on the prioritization of "authoritative sources." Indeed, such a method makes it particularly difficult for new brands to emerge, and it can affect the plurality of voices (Marsden and Meyer, 2019, p. 45). These concerns jibe with the ones exposed above regarding the importance of protecting freedom of expression and information by abstaining from defining the quality or reliability of content.

While all ethical implications regarding AI techniques developed to tackle disinformation online must be carefully considered, it is still important to bear in mind that human moderators need to cope with a very hard job. In May 2020, Facebook agreed to pay 52 million dollars to 11,250 moderators who had developed post-traumatic stress disorder from looking at the worst of the internet (The Economist, 2020b). Moreover, human review is prone to error or ambiguous results as well, and biases also affect the human mind. The right not to be subject to a decision solely based on automated processing.[34] is however relevant in this regard, and AI techniques used to tackle disinformation may put a publisher in this situation.

In parallel, the disinformation problem goes beyond search engines and platforms. Indeed, instant messaging platforms are also efficient means to spread disinformation "since they enable the sharing of content within closed groups with large numbers of users as well as the transfer of content from one closed group to another."[35] The murders in India mentioned above illustrate this issue. However, tackling disinformation in private messages would also raise significant concerns regarding the right to *privacy.*

[33] Equality before the law and the prohibition of discriminations are set out in articles 1 and 7 of the Universal Declaration of Human Rights, in article 14 of the European Convention on Human Rights and in article 21 of the Charter of Fundamental Rights of the European Union.

[34] This right is protected in the EU under article 22 of the GDPR.

[35] European Commission (2020a). The European Regulators Group for Audiovisual Media Services noted that "[e]ven though they may be defined as 'instant messaging tools' more than 'platforms', WhatsApp and Messenger allow the users to share content in closed groups that may contain an enormous amount of people, and offer simple functions to transfer messages/content from one group to another, thus making the content very easily viral. From this point of view, therefore, instant messaging tools may become very efficient tools to spread disinformation." ERGA (2020, p. 44, note 56).

4. The EU Regulation of Disinformation

The alarming consequences of the disinformation phenomenon online have prompted the EU institutions to initiate a regulatory framework to address the problem. That framework first represented a form of ascending (or "bottom-up") co-regulation, but it is now being reshaped into descending (or "top-down") co-regulation, predominantly through the Digital Services Act proposed by the Commission. While this new proposal focuses on the moderation of content circulating on online platforms, the EU parliament and the EU Council advocate for more positive measures aiming at enhancing media literacy and access to accurate information.

4.1. From ascending co-regulation …

Following a public consultation launched in 2017 revealing that over 97% of nearly 3,000 respondents were considering that they had already faced *fake news* (European Commission, 2018c), the European Commission set up in January 2018 a high-level group of experts (HLEG) to advise on policy initiatives to counter fake news and disinformation spread online. The report delivered by the HLEG (European Commission, 2018a). recommends the adoption of a controlled self-regulatory approach to address the problem. The European Commission Communication "Tackling online disinformation: a European approach" (European Commission, 2018d). does not deviate from this solution, and provides four principles and objectives that should guide action to tackle disinformation: "First, to improve *transparency* regarding the origin of information and the way it is produced, sponsored, disseminated, and targeted in order to enable citizens to assess the content they access online and to reveal possible attempts to manipulate opinion. Second, to promote *diversity* of information, in order to enable citizens to make informed decisions based on critical thinking, through support to high quality journalism, media literacy, and the rebalancing of the relation between information creators and distributors. Third, to foster *credibility* of information by providing an indication of its trustworthiness, notably with the help of trusted flaggers, and by improving traceability of information and authentication of influential information providers. Fourth, to fashion *inclusive* solutions. Effective long-term solutions require awareness-raising, more media literacy, broad stakeholder involvement and the cooperation of public authorities, online platforms, advertisers, trusted flaggers, journalists, and media groups" (European Commission, 2018d, p. 6).

In reaction to this EU position and following strictly the self-regulatory approach, representatives of the largest online platforms (Google, Facebook, Twitter, and Mozilla) and trade associations of advertisers presented a *Code of Practice on Disinformation.* The Commission, which had limited choice for alternative action, endorsed it in September 2018 (European Commission, 2018b). Microsoft and TikTok have also signed the code since.[36] It gathers commitments that the signatories can individually undertake and will address in good faith. These commitments relate to: 1° scrutiny of ad placements; 2° identification of political advertising and of issue-based advertising; 3° integrity of services through policies regarding bots and the use of automated systems; 4° empowerment of consumers by prioritizing authentic information, facilitating their evaluation of content via tools like indicators of trustworthiness, improving the findability of diverse perspectives about topics of public interest, and supporting efforts aimed at improving critical thinking and media literacy; 5° empowerment of research aimed at understanding the disinformation phenomenon and its impact. The *Code* also includes an annex containing best practices that signatories will apply to implement their commitments. The signatories further commit to publish annually a report on the different measures they undertake related to their specific commitments. This report must include, inter alia, indicators providing transparency regarding the scale of the disinformation problem, the submitted complaints, and the provided solutions.

The *Code of Practice on disinformation* represents a form of co-regulation that we name "ascending" since the initiative comes from private actors, the content has been decided by signatories and the execution

[36] Microsoft signed the *Code* in May 2019 and TikTok in June 2020.

is marginally controlled by public authorities through the review of the report by the Commission. In September 2018, the *Sounding Board of the Multistakeholder Forum on Disinformation*, which gathers representatives of the media, civil society, fact checkers and academia, had already expressed its skepticism regarding the "so-called *Code of Practice*" (Sounding Board, 2018). Its concerns are now mostly confirmed, as the recent assessment of the *Code* (European Commission, 2020a). exposes various shortcomings of such self-regulation: disparity between the reports and the measures undertaken by the signatories; lack of coverage of the sector; lack of participation of some key platforms (such as WhatsApp); lack of an independent oversight mechanism and of cooperation mechanisms; lack of access to data necessary to verify the signatories' practices; lack of consequences in case of breach; and lack of protection of fundamental rights, including through mechanisms for redress. This assessment therefore paves the way for a "descending" co-regulatory approach.

4.2. … Toward descending co-regulation

The Digital Services Act,[37] a new proposal for a Regulation presented in December 2020 by the Commission, proposes such a co-regulatory approach to tackle the disinformation problem, providing for a co-regulatory backstop for the measures that should be included in the revised and strengthened *Code of Practice on Disinformation.*[38] The revised and strengthened *Code of Practice* will build on the guidance of the Commission,[39] as announced in the European Democracy Action Plan (European Commission, 2020b). The declared purpose of the proposal is to enhance platforms' accountability by requiring them to assume their responsibility for the actions they take and the systemic risks they pose, in order to build a safe digital space where fundamental rights of all users of digital services are protected. The proposal provides for specific mechanisms and measures to tackle illegal content but does not define harmful content that is not illegal and does not subject it to removal obligations, while further prohibiting general monitoring obligations.[40] Nevertheless, and this is of particular interest for our study, some provisions relate to content moderation as such and to algorithmic systems that shape information flows online, which we specify here below.

Some transparency obligations are imposed on all providers of intermediary services[41]: they "shall include information on any restrictions that they impose in relation to the use of their service in respect of

[37] European Commission, Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (15 December 2020) 2020/0361 (COD).

[38] Digital Services Act, Article 35 and recitals 68 and 69.

[39] The guidance was published on May 26, 2021. European Commission (2021). "The Guidance calls for reinforcing the Code of Practice on Disinformation in the following areas to ensure a complete and consistent application across stakeholders and EU countries:

- Larger participation with tailored commitments.
- Better demonetising of disinformation.
- Ensuring the integrity of services.
- Improving the empowerment of users.
- Increasing the coverage of fact-checking and providing increased access to data to researchers.
- Creating a more robust monitoring framework.

The signatories of the Code of Practice should convene to strengthen the Code in line with the Commission's guidance and present a first draft in autumn."

We will analyse this new document in another contribution.

[40] Digital Services Act, Article 7.

[41] "'Intermediary service' means one of the following services:

- a 'mere conduit' service that consists of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network;
- a 'caching' service that consists of the transmission in a communication network of information provided by a recipient of the service, involving the automatic, intermediate and temporary storage of that information, for the sole purpose of making more efficient the information's onward transmission to other recipients upon their request;
- a 'hosting' service that consists of the storage of information provided by, and at the request of, a recipient of the service" Digital Services Act, Article 2(f).

information provided by the recipients of the service, in their terms and conditions. That information shall include information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review."[42] They also have to annually report on any content moderation they engage in.[43] The transparency reporting obligations are expanded for online platforms,[44] which must include in their report information on "any use made of automatic means for the purpose of content moderation, including a specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards applied;"[45] and very large online platforms[46] see their reporting obligations further expanded.[47]

Where a hosting service decides to remove or disable access to specific information provided by a recipient of the service, it must inform the recipient and provide a clear and specific statement of reasons. That statement contains, "where applicable, information on the use made of automated means in taking the decision, including where the decision was taken in respect of content detected or identified using automated means," but also information on the redress possibilities,[48] which should always include judicial redress.[49] These decisions and statement of reasons must also be published in a publicly available dataset managed by the Commission.[50] Following a decision taken by an online platform (to remove or disable access to information, to suspend or terminate the provision of the service to a recipient, or to suspend or terminate the recipient's account, on the ground that the information provided by the recipient is illegal or incompatible with their terms and conditions), internal complaint handling systems meeting certain conditions must be provided for,[51] but also possibility of out-of-court dispute settlement by independent certified bodies.[52]

Regarding advertising methods, online platforms have some basic transparency obligations. First, they "shall ensure that the recipients of the service can identify, for each specific advertisement displayed to each individual recipient, in a clear and unambiguous manner and in real time: (a) that the information displayed is an advertisement; (b) the natural or legal person on whose behalf the advertisement is displayed; and (c) meaningful information about the main parameters used to determine the recipient to whom the advertisement is displayed."[53] Second, in addition to those, very large online platforms are also required to ensure public access to repositories containing information on the advertisements they display on their online interface.[54] Such repositories are requested in order to "facilitate supervision and research into emerging risks brought about by the distribution of advertising online, for example in relation to […] manipulative techniques and disinformation with a real and foreseeable negative impact on public health, public security, civil discourse, political participation and equality."[55]

Very large online platforms (platforms providing their services to minimum 10% of the EU population) are also required to set out, in their terms and conditions, the main parameters used in their recommender systems (when they use such systems), in an easily comprehensible manner to ensure that the recipients

[42] Digital Services Act, Article 12.
[43] Digital Services Act, Article 13.
[44] "'Online platform' means a provider of a hosting service which, at the request of a recipient of the service, stores and disseminates to the public information, unless that activity is a minor and purely ancillary feature of another service and, for objective and technical reasons cannot be used without that other service, and the integration of the feature into the other service is not a means to circumvent the applicability of this Regulation." Digital Services Act, Article 2(h).
[45] Digital Services Act, Article 23 §1(c).
[46] Very large platforms are platforms providing their services to a number of average monthly active recipients of the service in the EU equal or higher than 10% of the EU population (currently amounting to 45 million). Digital Services Act, Article 25.
[47] Digital Services Act, Article 33.
[48] Digital Services Act, Article 15 §1–2.
[49] Digital Services Act, recital 42.
[50] Digital Services Act, Article 15 §4.
[51] Digital Services Act, Article 17.
[52] Digital Services Act, Article 18.
[53] Digital Services Act, Article 24.
[54] Digital Services Act, Article 30.
[55] Digital Services Act, Recital 63.

understand how information is prioritized for them, and to provide the recipients with the option to modify or influence such parameters, including at least one option that is not based on profiling.[56]

Furthermore, risk assessments and mitigation measures are required from very large online platforms. They are requested to identify, analyse and assess the significant systemic risks stemming from the functioning and use made of their services in the EU.[57] In this regard, some categories of risks must be assessed in-depth. Among these risks, the regulation proposal underlines the negative effects for the exercise of the fundamental rights, which "may arise […] in relation to the design of the algorithmic systems used by the very large online platforms, or the misuse of their service through […] methods for silencing speech or hampering competition;"[58] and, secondly, the intentional manipulation of their service, which "may arise […] through the creation of fake accounts, the use of bots, and other automated or partially automated behaviors, which may lead to the rapid and widespread dissemination of information that is illegal content or incompatible with an online platform's terms and conditions."[59] When conducting risk assessments, these very large online platforms are required to "take into account, in particular, how their content moderation systems, recommender systems and systems for selecting and displaying advertisement influence any of the systemic risks […], including the potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions."[60] They must then take measures to mitigate the identified risks, such as adapting content moderation or recommender systems.[61]

In order to ensure compliance of very large online platforms with their obligations laid down in the regulation, such platforms have to appoint compliance officers,[62] and provide the Digital Services Coordinator[63] of the establishment or the Commission with access to or reporting of specific data upon request.[64] An independent audit is also foreseen, only for very large online platforms,[65] "given the need to ensure verification by independent experts."[66]

We finally outline that the proposal allows the Commission to initiate the drawing up of crisis protocols "for addressing crisis situations strictly limited to extraordinary circumstances affecting public security or public health,"[67] in order to "coordinate a rapid, collective and cross-border response in the online environment."[68] It specifies that "[e]xtraordinary circumstances may entail any unforeseeable event, such as earthquakes, hurricanes, pandemics and other serious cross-border threats to public health, war and acts of terrorism, where, for example, online platforms may be misused for the rapid spread of illegal content or disinformation or where the need arises for rapid dissemination of reliable information."[69]

4.3. The Commission's omissions regarding important issues

From our analysis of the regulation proposal, we observe three points to which the Commission should have given more attention.

First, the proposal does not differentiate sufficiently the disinformation problem from the moderation of illegal content. The adopted approach regarding harmful content that is not *per se* illegal is questionable: while there was an agreement amongst stakeholders to not define such content and to not subject it to removal obligations "as this is a delicate area with severe implications for the protection of freedom of

[56] Digital Services Act, Article 29 and Recital 62.
[57] Digital Services Act, Article 26 §1.
[58] Digital Services Act, Recital 57.
[59] Digital Services Act, Recital 57.
[60] Digital Services Act, Article 26 §2.
[61] Digital Services Act, Article 27.
[62] Digital Services Act, Article 32.
[63] See Digital Services Act, Articles 39 and 41.
[64] Digital Services Act, Article 31 and Recital 64.
[65] Digital Services Act, Article 28.
[66] Digital Services Act, Recital 60.
[67] Digital Services Act, Article 37.
[68] Digital Services Act, Recital 71.
[69] Digital Services Act, Recital 71.

expression,"[70] providers of intermediary services are allowed to decide what content to moderate as they can impose restrictions in relation to the use of their service, as long as they inscribe those in their terms and conditions.[71] Then, they have to respect all their other obligations provided for in the proposal's provisions we outlined just above, which apply in a nondiscriminatory manner to all content moderation irrespective of the content's type. As we have already mentioned, it is for instance because Trump would have violated Twitter's and Facebook's terms and conditions that his accounts were permanently suspended by the platforms.[72] Admittedly, the regulation proposal provides for various transparency measures and requests the setting up of redress possibilities, but is it admissible to let private companies decide which content should be moderated, if we consider the overarching importance of freedom of expression? Facebook itself would prefer to make decisions about whether content is harmful "according to frameworks agreed by democratically accountable lawmakers."[73] In this context, we wish to clarify that false, inaccurate or misleading content is not the main element of the disinformation problem. What is truly and legally problematic is the manipulation of individuals' opinions, which is achieved cogently by leveraging the current digital ecosystem, as we demonstrated in Section 2.

Second, the proposal provides for enhanced responsibilities for very large online platforms given their importance "due to their reach, in particular as expressed in number of recipients of the service, in facilitating public debate, economic transactions and the dissemination of information, opinions and ideas and in influencing how recipients obtain and communicate information online,"[74] and because such platforms "may cause societal risks, different in scope and impact from those caused by smaller platforms."[75] As indicated above, the proposal specifies that such risks may stem from the use of algorithmic systems, the creation of fake accounts, or the use of automated behaviors such as bots.[76] In this context, it appears that such platforms are providing an essential service: they permit communication and reception of information online. Such essential service could be considered as a public or, at least, universal service like the service given by telecommunications' operators. Indeed, the proposal further acknowledges that such platforms "are used in a way that strongly influences safety online, the shaping of public opinion and discourse, as well as on online trade. The way they design their services is generally optimized to benefit their often advertising-driven business models and can cause societal concerns. In the absence of effective regulation and enforcement, they can set the rules of the game, without effectively identifying and mitigating the risks and the societal and economic harm they can cause."[77]

Third, we wish to underline that the Commission's self-conferred possibility to initiate the drawing up of crisis protocols in case of emergencies may be particularly problematic regarding freedom of expression and freedom of information, if applied. This provision has certainly been included in reaction to the many problems related to the infodemic stemming from the Covid-19. The infodemic is sure enough problematic, but what generates this infodemic? The digital ecosystem, again, as we explained in Section 2.2. Certainly, ensuring that reliable information is always available and even more during a time of crisis is legitimate, but content provided by the government should not silence people's voices, and we fear that the enforcement of crisis protocols would have such a consequence. The right approach would be to tackle the misuse of online platforms to spread disinformation, such misuse being furthermore one of the reasons for the Commission to include this tendentious provision.[78] It is however relevant to note that tackling such misuse and providing people with reliable information are always necessary, not only in times of crisis.

[70] Digital Services Act, p. 9.
[71] Digital Services Act, Article 12.
[72] Twitter, "Permanent suspension of @realDonaldTrump" (note 27); Facebook, "Referring Former President Trump's Suspension From Facebook to the Oversight Board" (note 27).
[73] Facebook, "Referring Former President Trump's Suspension From Facebook to the Oversight Board" (note 27).
[74] Digital Services Act, Recital 53.
[75] Digital Services Act, Recital 54.
[76] Digital Services Act, Recital 57.
[77] Digital Services Act, Recital 56.
[78] Digital Services Act, Recital 71.

4.4. A more positive approach by the EU Parliament and the Council of the EU

On its part, the European Parliament adopted in November 2020 a resolution on strengthening media freedom: protection of journalists in Europe, hate speech, disinformation and the role of platforms (European Parliament, 2020). This resolution focuses precisely on the right to freedom of expression and information and on democracy (European Parliament, 2020, para A), therefore adopting a more adequate approach to tackle the disinformation problem. It first draws attention to the essential role of the media and investigative journalists. Accordingly, the Resolution underlines the need to preserve their freedom and pluralism by, inter alia, maintaining their independence from political or governmental interference (European Parliament, 2020, paras 6–7); promoting measures aimed at financing and supporting media and independent journalism (European Parliament, 2020, paras 3, 21); protecting journalists and media workers (European Parliament, 2020, paras 10–12); avoiding media ownership concentration (European Parliament, 2020, para 16); and supporting the development of a vibrant and pluralistic media landscape through a EU media action plan (European Parliament, 2020, para 20). Regarding the role of platforms in countering disinformation, it emphasizes the requirement to avoid any drift to monopoly or concentration of information sources and over-censorship or removal (European Parliament, 2020, paras 35, 37); it promotes collaboration between fact checkers, academic researchers and stakeholders to identify, analyse and expose potential disinformation threats (European Parliament, 2020, para 40); it pleads for joint action to counter disinformation and underlines the key role that platforms should play to that end (European Parliament, 2020, para 41) in a transparent and accountable manner (European Parliament, 2020, para 42). Finally, the document stresses that "using automated tools in content moderation may endanger freedom of expression and information" (European Parliament, 2020, para 43), and, therefore, encourages the creation of tools to enable users to report and flag potential disinformation, and the review by independent and impartial third-party fact-checking organizations (European Parliament, 2020, para 45). Meanwhile, it also highlights, in line with our statements above, that "online platforms are part of the online public space in which public debate take place" (European Parliament, 2020, para 35); and that the business models based on micro-targeting advertising may generate negative impacts (European Parliament, 2020, para 39). The resolution also calls for measures to increase effectively media literacy (European Parliament, 2020, paras 46–49), which is essential as regards the growing flow of information online.

This resolution thus points out many important issues and promotes measures that are essential to enhance freedom of expression and information. Of particular value in light of our study are also the Parliamentary call on the Commission "to engage further with digital platforms [regarding the right of individuals not to be subject to pervasive online tracking across sites and applications] and to step up efforts to enforce the prohibition of such practices, combat the strategic, automated amplification of disinformation through the use of bots and fake profiles online, and increase transparency with respect to the financing and distribution of online advertising," and its call on all online platforms "to ensure that the algorithms that underpin their search functions are not primarily based on advertising" (European Parliament, 2020, para 39).

The EU Council conclusions on safeguarding a free and pluralistic media system (Council of the European Union, 2020), also adopted in November 2020, add some primordial points on sustainability, pluralism and trustworthiness, three of which we underline here as they were not already present is such clear terms elsewhere, and may address some of the main issues we pointed out regarding the spread of disinformation through the digital ecosystem. First, in order to enhance the user's autonomy, the Council concludes that "offers of personalized content should be based on criteria which have been provided voluntarily and/or selected by the user" (Council of the European Union, 2020, para 21). Second, it invites the Commission to "prevent public harm by addressing the manipulative dissemination techniques of disinformation" (Council of the European Union, 2020, para 42). Third and as already cited, it agrees that "with regard to the importance of freedom of speech, states and administrative regulatory authorities as well as private platform providers should abstain from defining quality content or the reliability of content itself" (Council of the European Union, 2020, para 39).

5. Conclusion

Though disinformation is a long-standing problem, AI systems present in the current digital ecosystem of the web participate in the problem's aggravation predominantly in two ways. First, they can be leveraged by malicious stakeholders in order to manipulate individuals in a particularly effective manner and at a huge scale. Secondly, they directly amplify the spread of such content. These AI systems are programmed to enhance engagement, and therefore the main contributing factor to the spread of disinformation is the business model of the web. Besides that, social media bots are widespread, and fake content is increasingly realistic.

In reaction to all negative effects of disinformation on society, and particularly in the context of the Covid-19 pandemic, social media platforms and search engines are increasingly requested to act against the spread of disinformation online. As a consequence, many AI systems are developed to tackle the problem. Yet, the use of AI systems to tackle disinformation content, the dissemination of which is amplified by other AI systems, is not a miraculous solution and would generate additional concerns.[79] Manipulative means should be impaired since they endanger many ethical values commonly shared at the international level, but in doing so, the risks incurred by ethical values and fundament rights, especially by freedom of expression and freedom to receive information, need particular attention.

As there is a clear tension between moderation of content online and free speech, the use of AI systems for such moderation would be particularly problematic. The European Commission's initiatives lack some perspectives in this regard, as they provide for transparency and redress measures but do not provide for any definition of harmful content that should be moderated by online platforms. Abstaining from defining such content is partially in line with the EU Council's conclusions, which properly outline that public regulatory authorities should not define the quality of content or its reliability, but platforms neither (Council of the European Union, 2020, para 39). Yet, while public authorities renounce any censorship, private platforms do not.

In our opinion, a more constructive approach would be to change the current digital ecosystem in light of the problems it generates. AI systems present in this ecosystem should indeed be adapted toward the respect of fundamental rights and ethical values. Adapting the business model of the Web would be an integral part of this process because as long as platforms and search engines count on the remuneration received from advertisers, engagement of the users may be aimed. Moving to a subscriber-based or pay-to-pay model for media and communication may not be the solution to advocate for, but the platforms' tendency to constantly increase revenues from advertisers by seeking users' engagement is disproportionate as regards the consequences of such practices. The obligation, included in the regulation proposal, for very large online platforms to provide their users with the possibility to modify or influence the parameters used in the platform's recommender systems, including at least one option which is not based on profiling,[80] is particularly relevant in this regard. The Draft Recommendation of the Council of Europe on the protection of individuals with regard to the processing of personal data in the context of profiling[81] goes along the same line, requesting from online intermediary services to "give data subjects both the possibility to opt in as regards the profiling and the choice between the different profiling purposes or degrees."

In order to enhance access to accurate information while respecting freedom of expression and information, support for media and journalists is needed, as the EU Parliament claims it. Media pluralism is indeed of particular importance. The establishment of a public information service providing citizens with controlled information can also be envisaged, but should by no means hamper the individuals' capacity to share their opinions or the media pluralism to be effective. We add that

[79] Furthermore, as noted by Mireille Hildebrandt, assuming that such systems can do without the acuity of human judgement would amount to "mistaking the imitation for what is imitated." Hildebrandt (2018).

[80] Digital Services Act, Article 29 and Recital 62.

[81] Council of Europe (Directorate General of Human Rights and Rule of Law), Consultative Committee of the Convention for the protection on individuals with regard to automatic processing of personal data—Convention 108, "Draft Recommendation on the protection of individuals with regard to the processing of personal data in the context of profiling (revising Recommendation (2010) 13)" (19 February 2021).

while content moderation through AI systems is not the right approach, such systems may still be used to counter the manipulation of the digital ecosystem, including through the accurate detection of AI-generated content, of social media bots, or through the detection of the malicious use of micro-targeting systems to target disinformation content at each user. It is indeed not the content itself that must be tackled, but rather the malicious use of technologies to amplify the conveyed message. Besides all these measures, media literacy is primordial. Indeed, while education to the critical evaluation of information has always been essential, it has substantially gained in importance with the growing flow of information permitted by our data-driven world. Freedom of expression, freedom of information, media pluralism, media literacy… all these are needed for democracy's preservation—or, may we say, for its revival.

Acknowledgments. A preprint version of this article was made available on ResearchGate. DOI: 10.13140/RG.2.2.28805.27365. We thank the reviewers from Data and Policy for the valuable comments they made on that first version.

Data Availability Statement. All resources used are included in references. Where they are available online, a link is provided.

Author Contributions. Both authors have contributed to the conceptualization, data curation, formal analysis, methodology, project administration, visualization, writing—original draft, writing—review and editing, and approved the final submitted draft.

Funding Statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing Interests. The authors declare no competing interests exist.

References

Akers J, **Bansal G**, **Cadamuro G**, **Chen C**, **Chen Q**, **Lin L**, **Mulcaire P**, **Nandakumar R**, **Rockett M**, **Simko L**, **Toman J**, **Wu T**, **Zeng E**, **Zorn B and Roesner F** (2018) Technology-Enabled Disinformation: Summary, Lessons, and Recommendations, Technical Report UW-CSE, 21 December 2018. Available at https://arxiv.org/abs/1812.09383 (accessed 01 March 2021).

Assenmacher D, **Clever L**, **Frischlich L**, **Quandt T**, **Trautmann H and Grimme C** (2020) Demystifying social bots: On the intelligence of automated social media actors. *Social Media & Society 6*(3), p. 1. https://doi.org/10.1177/2056305120939264

Balestrucci A (2020) How Many Bots Are You Following? 15 January 2020. Available at https://arxiv.org/abs/2001.05222v1 (accessed 01 March 2021).

Bergamini D (2020) Need for Democratic Governance of Artificial Intelligence. Committee on Political Affairs and Democracy – Council of Europe, 24 September 2020, no. 15150. Available at https://pace.coe.int/en/files/27616 (accessed 01 March 2021).

Bindner L and Gluck R (2020) Social Media and the Murder of Samuel Paty, Global Network on Extremism and Terrorism, 6 November 2020. Available at https://gnet-research.org/2020/11/06/social-media-and-the-murder-of-samuel-paty/ (accessed 01 March 2021).

Council of the European Union (2020) Council Conclusions on Safeguarding a Free and Pluralistic Media System, 13260/20, Brussels, 27 November 2020. Available at https://data.consilium.europa.eu/doc/document/ST-13260-2020-INIT/en/pdf (accessed 01 March 2021).

de Coorebyter V (2020) L'Internet: démocratie ou démagogie. In Poullet Y (ed.), *Vie privée, liberté d'expression et démocratie dans la société du numérique*. Brussels: Larcier, p. 249.

Dixit P and Mac R (2018) How WhatsApp Destroyed a Village. *Buzzfeed News*, September 2018. Available at https://www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india (accessed 01 March 2021).

EDPS Ethics Advisory Group (2018) Towards a Digital Ethics – Report. Available at https://edps.europa.eu/data-protection/our-work/publications/ethical-framework/ethics-advisory-group-report-2018_en (accessed 01 March 2021).

ERGA (2020) *ERGA Report on disinformation: Assessment of the implementation of the Code of Practice* (2020). Available at https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf (accessed 01 March 2021).

European Commission (2018a) *A Multi-dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation.* Directorate-General for Communication Networks, Content and Technology. Available at https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation (accessed 01 March 2021).

European Commission (2018b) *Code of Practice on Disinformation*. Available at https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

European Commission (2018c) Synopsis Report of the Public Consultation on Fake News and Online Disinformation. Available at https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation (accessed 01 March 2021).

European Commission (2018d) *Tackling Online Disinformation: A European Approach (Communication) COM(2018) 236 final*. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236 (accessed 01 March 2021).

European Commission (2020a) *Assessment of the Code of Practice on Disinformation — Achievements and areas for further improvement*. Commission Staff working document (SWD(2020) 180 final).

European Commission (2020b) *European Democracy Action Plan (Communication) COM(2020) 790 final.* Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423 (accessed 01 March 2021).

European Commission (2021) Guidance on Strengthening the Code of Practice on Disinformation (COM(2021) 262 final). Available at https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation (accessed 03 July 2021).

European Parliament (2020) European Parliament Resolution on Strengthening Media Freedom: The Protection of Journalists in Europe, Hate Speech, Disinformation and the Role of Platforms (25 November 2020) 2020/2009(INI). Available at https://www.europarl.europa.eu/doceo/document/TA-9-2020-0320_EN.html (accessed 01 March 2021).

Festré A and Garrouste P (2015) The 'Economics of attention': A history of economic thought perspective. *Œconomia. History, Methodology, Philosophy 5*(1), 3–36. https://doi.org/10.4000/oeconomia.1139

Fink C (2018) Dangerous speech, anti-Muslim violence and Facebook in Myanmar, Special issue: Contentious narratives: Digital technology and the attack on liberal democratic norms. *Journal of International Affairs 71*(1.5), pp. 43–52. Available at https://jia.sipa.columbia.edu/dangerous-speech-anti-muslim-violence-and-facebook-myanmar (accessed 01 March 2021).

Hannah M (2021) QAnon and the information dark age. *First Monday 26*(2). https://doi.org/10.5210/fm.v26i2.10868

Hanot M and Michel A (2020) Entre menaces pour la vie en société et risques réglementaires, les *fake news*: un danger pour la démocratie? In Poullet Y (ed.), *Vie privée, liberté d'expression et démocratie dans la société du numérique*. Brussels: Larcier .

Harbinja E and Karagiannopoulos V (2019) Web 3.0: The Decentralised Web Promises to Make the Internet Free Again. *The Conversation*, 11 March 2019. Available at https://theconversation.com/web-3-0-the-decentralised-web-promises-to-make-the-internet-free-again-113139 (accessed 01 March 2021).

Hildebrandt M (2018) Primitives of Legal Protection in the Era of Data-Driven Platforms, March 2018, p. 11. http://dx.doi.org/10.2139/ssrn.3140594

Howard PN and Kollanyi B (2016) Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. http://dx.doi.org/10.2139/ssrn.2798311

Human Rights Council (2018) Report of the Independent International Fact-Finding Mission on Myanmar, A/HRC/39/64, 12 September 2018. Available at https://www.ohchr.org/en/hrbodies/hrc/myanmarFFM/Pages/ReportoftheMyanmarFFM.aspx (accessed 01 March 2021).

Kertysova K (2018) Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights 29*, 55–81. https://doi.org/10.1163/18750230-02901005

Kirby EJ (2016) The City Getting Rich from Fake News. *BBC News*, 5 December 2016. Available at https://www.bbc.com/news/magazine-38168281 (accessed 01 March 2021).

Lamo M and Calo R (2018) *Regulating Bot Speech.* UCLA Law Review, 16 July 2018, p. 1. http://dx.doi.org/10.2139/ssrn.3214572

Lance Bennett W and Livingston S (eds) (2020) *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108914628

Maréchal N and Biddle ER (2020) It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge - A Report from Ranking Digital Rights, New America, 17 March 2020. Available at https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/ (accessed 01 March 2021).

Marsden C and Meyer T (2019) Regulating Disinformation with Artificial Intelligence: Effect of Disinformation Initiatives on Freedom of Expression and Media Pluralism. European Parliamentary Research Service (EPRS) – Scientific Foresight Unit (STOA), March 2019, PE 624.279. Available at https://op.europa.eu/en/publication-detail/-/publication/b8722bec-81be-11e9-9f05-01aa75ed71a1 (accessed 01 March 2021).

Mork A (ed.) (2020) *Fake for Real – A History of Forgery and Falsification*. House of European History, Temporary Exhibition Catalogue. Luxembourg: Publications Office of the European Union.

Newman N, **Fletcher R**, **Schulz A**, **Andı S and Nielsen RK** (2020) Reuters Institute Digital News Report 2020. Available at https://www.digitalnewsreport.org/ (accessed 01 March 2021).

Partnership on AI (2020) *The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity.* Available at https://www.partnershiponai.org/ai-and-media-integrity-steering-committee/.

Perrigo B (2019) The World Wide Web Turns 30 Today. Here's How Its Inventor Thinks We Can Fix It. *Time*, 12 March 2019. Available at https://time.com/5549635/tim-berners-lee-interview-web/ (accessed 01 March 2021).

Poullet Y (2020a) *La "Révolution" numérique: quelle place encore pour le droit?* Brussels: Académie royale de Belgique, p. 78.

Poullet Y (2020b) *Ethique et droits de l'Homme dans notre société du numérique*. Brussels: Académie Royale de Belgique.

Poullet Y (2020c) Vue de Bruxelles. Modes alternatifs de régulation et libertés dans la société du numérique. In Castets-Renard C, Ndior V and Rass-Masson L (eds), *Enjeux internationaux des activités numériques: entre logique territoriale des Etats et puissance des acteurs privés*. Brussels: Larcier, p. 108.

Rundle M and Conley C (2007) Ethical Implications of Emerging Technologies: A Survey. UNESCO – Information for All Programme. Available at https://unesdoc.unesco.org/ark:/48223/pf0000149992 (accessed 01 March 2021).

Sankin A (2020) Want to Find a Misinformed Public? Facebook's Already Done It. The Markup, 23 April 2020. Available at https://themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it (accessed 01 March 2021).

Sounding Board (2018) *The Sounding Board's unanimous final opinion on the co-called Code of Practice*. Available at https://www.ebu.ch/files/live/sites/ebu/files/News/2018/09/Opinion%20of%20the%20Sounding%20Board.pdf (accessed 01 March 2021).

The Economist (2020a) Disinformation in Myanmar – Anti-social Network, Who Controls the Conversation? – Social Media and Free Speech, 24 October 2020, pp. 45–46.

The Economist (2020b) The Great Clean-Up, Who Controls the Conversation? – Social Media and Free Speech, 24 October 2020, p. 19.

Varol O, **Ferrara E**, **Davis CA**, **Menczer F and Flammini A** (2017) Online Human-Bot Interactions: Detection, Estimation, and Characterization. Available at https://arxiv.org/abs/1703.03107 (accessed 01 March 2021).

Walorska AM (2020) *Deepfakes and Disinformation.* Friedrich Naumann Foundation for Freedom, May 2020. Available at https://fnf-europe.org/wp-content/uploads/2020/06/fnf_deepfakes_broschuere_en_web.pdf (accessed 01 March 2021).

World Health Organization (WHO) (2020a) *Coronavirus Disease 2019 (COVID-19) Situation Report-45*. Available at https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200305-sitrep-45-covid-19.pdf?sfvrsn=ed2ba78b_4.

World Health Organization (WHO) (2020b) The Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC, *Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation*. Available at https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

Wright R (2001) The Man Who Invented the Web. *Time*, 24 June 2001. Available at http://content.time.com/time/magazine/article/0,9171,137689,00.html (accessed 01 March 2021).

Zago M, **Nespoli P**, **Papamartzivanos D**, **Perez MG**, **Marmol FG**, **Kambourakis G and Perez GM** (2019) Screening out social bots interference: Are there any silver bullets? *IEEE Communications Magazine 57*(8), 98. https://doi.org/10.1109/MCOM.2019.1800520

Cite this article: Bontridder N and Poullet Y (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3: e32. doi:10.1017/dap.2021.20